9/3/2017.

Στατιστικός

Πληθυσμός ← Τυχαίο Δείγμα
(Αντιπροσωπευτικό υποσύνολο του πληθυσμού)

Έστω n (μέγεθος του δείγματος) μέλη του πληθυσμού απαρτίζουν το τυχαίο δείγμα (τ.δ.)

Μετρώ το χαρακτηριστικό X στα n-μέλη του δείγματος (τ.δ.).
Προκύπτει έτσι το τ.δ. $x_1, x_2, \dots, x_n$

κατανομή των δεδομένων

$\bar{x} = \frac{1}{n}\sum x_i$

$s^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2$

Ένα ή περισσότερα χαρακτηριστικά γνωρίσματα των μελών του πληθυσμού.

π.χ. τυχαία μεταβλητή X ←

Γνωρίζω ότι ∃ η κατανομή της X

Θα ασχοληθούμε

αν γνωρίζω τη συναρτησιακή έκφραση π.χ. f_x (πυκνότητα)

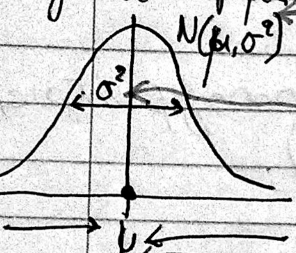
↑ Παραμετρική στατιστική

αν δεν γνωρίζω τη συναρτησιακή έκφραση π.χ. f_x

↑ Μη παραμετρική στατιστική

Έστω $X \sim N(\mu, \sigma^2)$
άγνωστες παράμετροι

$N(\mu, \sigma^2)$



(αν το δείγμα είναι αντιπροσωπευτικό!!!)

Πρόβλημα: Η προσέγγιση ή η εκτίμηση των άγνωστων παραμέτρων της κατανομής

Στατιστική Συμπερασματολογία:
Κλάδος στατιστικής που αναπτύσσει μεθοδολογίες για την προσέγγιση (εκτίμηση) των άγνωστων παραμέτρων της κατανομής που περιγράφει την τ.μ. X

I Εκτίμηση σε σημείο

II Εκτίμηση σε διάστημα (διαστήματα εμπιστοσύνης)

III Έλεγχοι Στατιστικών Υποθέσεων (Στατιστικά τεστ)

Εκτίμηση σε σημείο

Συμβολισμός - Ορολογία

• Έστω τυχαίο δείγμα (τ.δ.) μεγέθους n από πληθυσμό με κατανομή $f(x,\theta)$ η οποία εξαρτάται από μια παράμετρο $\theta \in \Theta \subseteq \mathbb{R}$ (ή $\mathbb{R}^n$)

Ειδικότερα το τδ $x_1, \dots, x_n$ υπακούει σε δύο απαιτήσεις:

(i) Οι $x_1, \dots, x_n$ είναι τ.μ. ανεξάρτητες

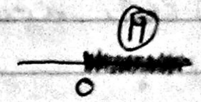
(ii) Οι $x_1, \dots, x_n$ είναι ισόνομες, δηλ. περιγράφονται από την ίδια κατανομή $f(x,\theta)$

Το σύνολο Θ ← Παραμετρικός χώρος
και είναι ανοικτό $\subseteq \mathbb{R}$

π.χ. Έστω τ.δ. $x_1, \dots, x_n$ από πληθυσμό με $\text{Εκθ}(\theta)$

$$f(x,\theta) = \theta e^{-\theta x}, \quad x>0, \ \theta>0$$

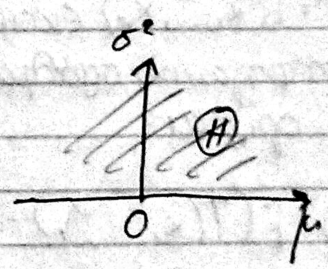
Άρα $\Theta = \{\theta \in \mathbb{R} : \theta > 0\} = (0,\infty)$



Έστω τ.δ. $x_1, \dots, x_n$ από πληθυσμό $N(\mu,\sigma^2)$, $f(x,\theta) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

$-\infty < \mu < \infty$, $\sigma^2 > 0$

$$\Theta = \left\{ \theta = (\mu,\sigma^2) : \begin{matrix} \mu \in \mathbb{R} \\ \sigma^2 > 0 \end{matrix} \right\}$$



Στατιστική Συνάρτηση

Έστω τ.δ. $x_1, \dots, x_n$ από πληθυσμό με κατανομή $f(x,\theta)$, $\theta \in \Theta$

Στατιστική Συνάρτηση: Κάθε συνάρτηση του τ.δ. $x_1, \dots, x_n$ και $T = T(X_1, \dots, X_n)$

π.χ. στατιστικών συναρτήσεων: $T = \bar{x} = \frac{1}{n} \sum x_i$

$$T = S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

Εκτιμητής: Τη στατιστική συνάρτηση που χρησιμοποιούμε για την εκτίμηση (προσέγγιση) της άγνωστης παραμέτρου θ ή μιας συνάρτησης $g(\theta)$ της θ

Κριτήρια Επιλογής Εκτιμητών - Μέσο τετραγωνικό σφάλμα

Έστω ένας εκτιμητής $T = T(x_1, \ldots, x_n)$ της $g(\theta)$, $\theta \in \Theta$
Ο T θα είναι καλός εκτιμητής της $g(\theta)$ αν $|T - g(\theta)|$ είναι μικρό

Αδυναμίες $|T-g(\theta)|$

(i) Η απόλυτη τιμή είναι δύσχρηστη
(ii) Η $|T-g(\theta)|$ εξαρτάται από το τ.δ. $x_1, \ldots, x_n$
Για να αρθούν οι αδυναμίες (i) και (ii)
Ο T θα είναι καλός εκτιμητής αν η τιμή της $E[T-g(\theta)]^2$ είναι μικρή

Ορισμός: (Μέσο τετραγωνικό Σφάλμα)
Έστω $T = T(x_1, \ldots, x_n)$ εκτιμητής της $g(\theta)$
Το μέσο τετραγωνικό σφάλμα του εκτιμητή T για την εκτίμηση της $g(\theta)$ ορίζεται:

$$ΜΤΣ(T, \theta) = E(T(x_1, \ldots, x_n) - g(\theta))^2 = E(T - g(\theta))^2$$

Πρόταση: $ΜΤΣ(T, g(\theta)) = Var(T) + (E(T) - g(\theta))^2$

Απόδειξη: $Var(W) = E(W^2) - (EW)^2$

$ΜΤΣ(T, g(\theta)) \overset{\text{ορ}}{=} E(T - g(\theta))^2$

$= Var(T - g(\theta)) + (E(T - g(\theta)))^2$

$= Var(T) + (E(T) - g(\theta))^2$

$Var\, g(\theta) = 0$ αφού $g(\theta)$ σταθερά
$Var(T - g(\theta)) = Var T + Var\, g(\theta)$

Έστω τ.δ. $X_1,\dots,X_n$ από πληθυσμό με κατανομή $Exp(1/\theta)$, $\theta>0$.
Έστω ο εκτιμητής $\bar{X}$
Να βρεθεί ΜΤΣ $(\bar{X},\theta)$

Από πρόταση $ΜΤΣ(\bar{x},\theta)=Var(\bar{x})+(E(\bar{x})-\theta)^2$

$E(\bar{x})=E(\frac{1}{n}\sum X_i)=\frac{1}{n}\sum E(X_i)=\frac{1}{n}\sum\theta=\theta$

$Var(\bar{x})=Var(\frac{1}{n}\sum X_i)=\frac{1}{n^2}Var(\sum X_i)$

$$Var\left(\sum_{i=1}^{n}a_i w_i\right)=\sum_{i=1}^{n}a_i^2 Var(w_i)+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}a_i a_j Cov(w_i,w_j)$$

Αν W_i ανεξαρ. τότε $Cov(W_i,W_j)=0$ και έτσι $Var(\sum_{i=1}^{n}a_i w_i)=$
$=\sum_{i=1}^{n}a_i^2 Var(w_i)$

$$\xrightarrow[\text{ανεξαρτ.}]{X_i}\quad Var(\bar{x})=\frac{1}{n^2}\sum_{i=1}^{n}Var(x_i)=\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{(1/\theta)^2}=\frac{1}{n^2}\sum_{i=1}^{n}\theta^2$$

$$\Rightarrow Var(\bar{x})=\frac{\theta^2}{n}$$

Άρα $\boxed{ΜΤΣ(\bar{x},\theta)=\frac{\theta^2}{n}+0^2=\frac{\theta^2}{n}}$

Παράδειγμα: Αν $n\to\infty$ τότε $ΜΤΣ(\bar{x},\theta)\to 0$

Ορισμός: Έστω οι εκτιμητές T_1 και T_2 για την εκτίμηση της $g(\theta)$
Ο T_1 θα λέγεται καλύτερος από τον T_2 με κριτήριο το ΜΤΣ αν:
(α) $ΜΤΣ(T_1,\theta)\le ΜΤΣ(T_2,\theta)$, $\forall\theta\in\Theta$ παραμετρικός χώρος
(β) $ΜΤΣ(T_1,\theta_0)<ΜΤΣ(T_2,\theta_0)$, για κάποιο $\theta_0\in\Theta$

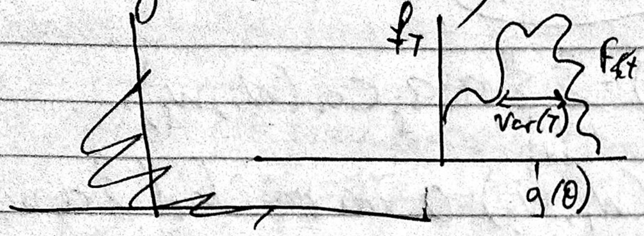
Παρατήρηση: (α) Ο T_2 θα λέγεται μη-αποδεκτός
(β) Ένας εκτιμητής λέγεται αποδεκτός αν δεν υπάρχει καλύτερος από αυτόν.

Καλύτερος εκτιμητής είναι εκείνος που έχει το μικρότερο ΜΤΣ
Αλλά αποδείξαμε ότι

$$ΜΤΣ(T,\theta) = Var(T) + (E(T) - g(\theta))^2$$

→ Καλύτερος εκτιμητής είναι εκείνος που (i) έχει τη μικρότερη διακύμανση ($Var(T)$)
(ii) τη μικρότερη $(E(T) - g(\theta))^2$.

Είναι λογικά τα (i) και (ii);



Ορισμός (Αμεροληψίας)

Ο εκτιμητής $T = T(X_1, \dots, X_n)$ λέγεται αμερόληπτος της $g(\theta)$ αν
$E(T) = g(\theta)$.

Ορισμός (Μεροληψίας)

Η μεροληψία του εκτιμητή T της $g(\theta)$ συμβολίζεται με $b(T,\theta)$ και είναι
$b(T,\theta) = E(T) - g(\theta)$

Πόρισμα: (α) $ΜΤΣ(T,\theta) = Var(T) - b^2(T,\theta)$
(β) Αν ο T αμερόληπτος της $g(\theta)$ τότε $b(T,\theta) = 0$
(γ) Αν ο T αμερόληπτος της $g(\theta)$ τότε $ΜΤΣ(T,\theta) = Var(T)$

Το (γ) παρακινεί για τον ορισμό ΑΟΕΔ εκτιμητή

Ορισμός Ο εκτιμητής T της $g(\theta)$ λέγεται αμερόληπτος ομοιόμορφα ελάχιστης διακύμανσης (ΑΟΕΔ) αν ο T είναι αμερόληπτος της $g(\theta)$ και έχει τη μικρότερη διακύμανση από όλους τους άλλους αμερόληπτους εκτιμητές της $g(\theta)$

Παράδειγμα 1 Έστω τ.δ. $X_1,\dots,X_n$ από πληθυσμό με μέση τιμή μ και διακύμανση σ^2

ΝΔΟ (i) ο $\bar{X}$ αμερόληπτη εκ μ

(ii) $Var(\bar{X}) = \frac{\sigma^2}{n}$

(i) Αρκεί νδο $E(\bar{X}) = \mu$ (από τον ορισμό)

$$E(\bar{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum \mu = \frac{1}{n} n\mu = \mu$$ Για όλες τις κατανομές ισχύει.

(ii) $$Var(\bar{X}) = Var\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2} Var\left(\sum X_i\right) \overset{X_1,\dots,X_n \text{ ανεξ.}}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Άρα ο $\bar{X}$ είναι αμερόληπτος του πληθυσμού.

Παράδειγμα 2

Έστω τ.δ. $X_1,\dots,X_n$ από πληθυσμό με μέση τιμή μ και διακύμανση σ^2

ΝΔΟ η δειγματική διακύμανση

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ είναι αμερόληπτος εκτιμητής σ^2

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n-1}\sum(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \frac{1}{n-1}\left\{\sum X_i^2 - 2\bar{X}\overbrace{\sum X_i}^{n\bar{X}} + n\bar{X}^2\right\} =$$

$$= \frac{1}{n-1}\left(\sum X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right) \Rightarrow S^2 = \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right)$$

$$E(S^2) = E\left\{\frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right)\right\} = \frac{1}{n-1}\left\{\sum E(X_i^2) - nE(\bar{X}^2)\right\} =$$

$Var(W) = EW^2 - (EW)^2$

$$\frac{1}{n-1}\left\{\sum\left(Var(X_i) + (EX_i)^2\right) - n\left(Var(\bar{X}) + (E\bar{X})^2\right)\right\} =$$

$$= \frac{1}{n-1}\left\{\sum(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right\} = \frac{1}{n-1}\left\{n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2}\right\} = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$